Integrating geospatial and public health data: examples using National Center for Health Statistics (NCHS) data systems

Lauren Rossen, Patsy Lloyd

National Center for Health Statistics
3311 Toledo Road
Hyattsville MD 20782 USA

Abstract

Previous research has shown that health is related to where people live, work, and play. The National Center for Health Statistics (NCHS) collects data from birth and death records, medical records, and health surveys. Although the publicly available data files do not typically include geographic data, geocoded data from many of the NCHS data systems are accessible through the Research Data Center (RDC). Integrating geospatial data with NCHS data systems provides researchers the opportunity to examine how various social and environmental determinants of health relate to specific health outcomes, behaviors, risk factors or disparities. Additionally, the linkage of geospatial data can augment NCHS data systems by providing multi-level or temporal information that would otherwise be unavailable in cross-sectional surveys or vital statistics data.

The objective of this presentation is to illustrate how geospatial data can be used in public health research and highlight key considerations for researchers conducting geospatial analyses. We will present examples of research using geospatial data to examine a broad range of public health topics, including food insecurity and diet among children, and associations between air pollution and health outcomes. These examples will illustrate key considerations that researchers should be aware of when using geospatial data for public health research purposes, including: limitations related to confidentiality and disclosure risk; selection bias due to linkage refusal, or failure to geocode; measurement error or bias due to temporal inconsistencies and/or misalignment of geographic boundaries; and issues related to ecological fallacies. Additional methodological concerns may include multi-level modeling with complex survey data, combining multiple sources of uncertainty, and appropriate methods for smoothing.

Ultimately, integrating geospatial data with NCHS data systems enhances our ability to investigate broader social, economic and environmental determinants of health and disparities. Researchers should be aware of the various strengths and limitations of using geospatial data in their analyses.

1. Introduction

Where an individual lives, works, and plays has important implications for their health [1, 2]. The inequitable distribution of neighborhood hazards, environmental exposures or access to health promoting resources also has implications for health disparities [3-5]. For instance, common exposures, including pollution or high-levels traffic, may be concentrated in specific geographic areas and have been linked to adverse health outcomes such as asthma, hospitalizations and mortality [6-9]. Alternatively, specific features of built environments, such as the presence of supermarkets, parks, or walkable communities have been linked to health promoting activities such as fruit and vegetable intake or physical activity [10-12].

As the Nation's principal health statistics agency, the National Center for Health Statistics (NCHS) collects and reports on data on the health of the U.S population using birth and death records, medical records, and health surveys. Although the publicly available data files do not typically include geographic data, geocoded data including state, county, census tract, and longitude/latitude identifiers from many of the NCHS data systems are accessible through the Research Data Center (RDC). For vital records, the geographic data may represent the individual's residence at birth or death; and, for health surveys, these data are typically the survey participant's residence at the time of interview.

Integrating geospatial data with NCHS data systems provides researchers the opportunity to examine how various social and environmental determinants of health relate to specific health outcomes, behaviors, risk factors or disparities. Additionally, the linkage of geospatial data can augment NCHS data systems by providing multi-level or temporal information that would otherwise be unavailable in analyses of cross-sectional survey or vital statistics system data.

The objective of this presentation is to illustrate how geospatial data can be used in public health research using geocoded population health surveys and will highlight key considerations for researchers conducting geospatial analyses. We will describe two research examples using geospatial data from NCHS data systems and highlight important considerations for analyses. And, in the final section of the presentation we draw some conclusions based on previous examples and considerations.

2. Examples

Example 1. Food insecurity and Diet

In 2012, approximately 8.3 million children (11.3% of children in the U.S.) lived in households in which at least one child experienced food insecurity, a condition characterized by a lack of consistent, dependable access to enough food for active and healthy living [13]. This first example examines associations between child-level and household-level food insecurity and dietary intake patterns. This analysis incorporates geospatial data from a variety of sources to account for neighborhood- and county-level factors related to both the experience of food insecurity among children in the U.S. and dietary intake [13].

Data

Data were from the 2007-2010 National Health and Nutrition Examination Survey (NHANES), a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES collects detailed information obtained through questionnaires administered in the home, as well as standardized physical examinations and laboratory data obtained in mobile examination centers. Specific subgroups are oversampled, including Hispanic and non-Hispanic black persons, and low-income persons. NHANES follows a four-stage sample design: the first stage is the selection of primary sampling units (PSUs) from all U.S. counties, stratified by geography and minority density; the second stage samples area segments (e.g., census blocks or combinations); the third stage selects dwelling units or households; and the fourth stage selects persons within households resulting in a sample of approximately 10,000 participants every two years.[14]

The study population for this analysis consisted of children and adolescents (2-15 years) who completed at least one 24-hour dietary recall in the mobile examination center. Pregnant adolescents and breastfeeding children were excluded leaving a sample of 5183 children and adolescents, 5136 of whom had data on food security (99%). More details about NHANES can be located on the NHANES webpage: http://www.cdc.gov/nchs/nhanes.htm.

Methods

Dietary intake was assessed via 24-hour recall completed by a trained interviewer with a second recall conducted for most participants via telephone 3-10 days later. During the household interview, an adult respondent completed the U.S. Food Security Survey Scale and provided information about participation in a variety of food assistance programs, receipt of free or reduced price school lunch and/or breakfast, participation in the Women, Infant and Children's Nutrition (WIC) program (for children 0-5 years), whether the child had public or private health insurance, whether the child was foreign-born, caregiver marital status (i.e., single, married/cohabitating, divorced/separated/widowed), caregiver education level (i.e., less than High School, High School/GED, some college, college or higher) and age, household size, race/ethnicity, income-to-poverty ratio (PIR), and household smoking. Child-level food insecurity as well as household food insecurity were both examined. Results described below refer to household-level food insecurity; results for child-level food insecurity can be seen in Rossen & Kobernik [13].

Restricted data files with geographic identifiers were used to link participating children to auxiliary data from several sources (e.g., Area Resource File, U.S. decennial Census data, U.S.D.A. Food Environment Atlas; see [13] for a complete list of data sources) based on their census tract and county of residence. These data sources and covariates included: tract- and county-level socio-demographic and economic characteristics (e.g., racial and ethnic population distribution, population size, % poverty, median household income, % of residents with less than High School education); residential racial segregation; the number of arrests per 100,000 county residents; the urban-rural designation of each county; county-level food store density (e.g., the number of grocery stores, convenience stores, fast food restaurants), food price (e.g., the price of milk, soda, fruit), and food assistance participation rates.

Propensity score weighting was used to ensure food secure and food insecure children are otherwise comparable with respect to characteristics at the individual, family, and geographic area-level [15]. Inverse probability of treatment weights (IPTW) were created from the propensity scores, and combined with the day one dietary survey weights in order to provide unbiased estimates of the average treatment effect (ATE) that are generalizable to the survey's target population [15]. In this case, the ATE represents the effect of food insecurity and the target population is U.S. children aged 2-15 years.

Usual intake of various dietary components was modeled using methods developed by the National Cancer Institute [16]. Post-stratified balanced repeated replicate (BRR) weights were used to account for the complex survey design.

Results

Approximately 16% of children 2-15 years old experienced food insecurity (i.e., marginal, low, or very low food security) in the prior 12 months. There were substantial differences between food secure and food insecure children across several demographic and socioeconomic characteristics, as well as neighborhood or contextual covariates. Food insecure children were more likely to live in census tracts with high deprivation, as well as experience both individual-level poverty and neighborhood deprivation. Food insecure children also resided in smaller, more urban counties with higher levels of crime compared to food secure children.

After propensity-score weighting, food secure and insecure children were balanced with respect to included covariates. In unadjusted models without propensity score weighting, there were significant differences in dietary intake by household food security status. Children in households that were food secure consumed 0.10 fewer cups of juice (SE 0.04, P<0.05), 0.12 more servings of whole grains (SE 0.04, P<0.05), 1.82 fewer teaspoons of added sugar (SE 0.61, P<0.05), 42.19 fewer kcal from solid fats and added sugars (SE 18.89, P<0.05), and a greater variety of food items (0.85 more foods, SE 0.21, P<0.05) with lower average caloric density (by 12.34 kcal, SE 3.02, P<0.05) compared to children in food insecure households (consisting of households reporting moderate, low or very low food security). In IPTW models, all of these differences were completely attenuated, suggesting that the set of confounders included in the estimation of the propensity score weights account for observed crude/unadjusted differences in dietary intake by household food security status.

Analytic Considerations

In this analysis, the classification of children as food insecure or residing in food insecure households was based on responses to questions asked in relation to specific experiences or behaviors over the past 12 months. Thus, there might be some misclassification of children's current experience of food insecurity, which would be expected to be more closely related to their current dietary patterns. Similarly, there may also be misclassification of tract-level or county-level characteristics for a variety of reasons. This could arise from temporal misalignment, where data from the decennial census were used to characterize census tracts in 2007. Additionally, children may have moved, and there may be differences in the associations between historical neighborhood/area-level exposures or current exposures and health outcomes of interest. For some outcomes, early-life exposures may be critical, whereas for other outcomes, more proximal exposures might be more relevant. Modifiable areal unit problems (MAUP) may also be an issue, where patterns for larger units such as counties may be different than for smaller-scale resident-defined neighborhoods. This is particularly notable for certain covariates in this analysis such as features of the food environment, which are known to exhibit substantial variation within counties or cities. Given that this analyses included geospatial information in the estimation of propensity score weights, any residual confounding, measurement error or bias in the assessment of these covariates would result in a failure of the propensity score methods to account for imbalances between the food secure and food insecure groups.

Example 2. Air Pollution and Childhood Respiratory Allergies in the United States: An application of linking the National Health Interview Survey (NHIS) to the Air Quality System (AQS)

Our second example is from Parker et al., who evaluated the association between air pollution and childhood respiratory allergies using air monitoring data from the Environmental Protection Agency (EPA) linked to respondent data of the 1987-2005 NHIS [17].

Data

NHIS is an annual cross-sectional household interview survey administered by the National Center for Health Statistics (NCHS) and serves as the principal source of information on the health of the civilian noninstitutionalized population of the United States. The survey uses a two-stage sampling design with primary sampling units (PSU's) in the first stage covering the 50 States and the District of Columbia and a second stage of addresses selected from PSUs. More details about NHIS can be located on the NHIS webpage: http://www.cdc.gov/nchs/nhis/about_nhis.htm [18].

The Environmental Protection Agency (EPA) collects and reports data on ambient concentrations from several thousand monitoring locations throughout the U.S. via the Air Quality System (AQS). Annual arithmetic averages are calculated for several principal air pollutants and provided on their Annual Summary Web pages. Data are collected for regulatory purposes, and the collection times, monitor locations and the measures vary for different pollutants. More information about these data can be found on the AQS webpage: http://www2.epa.gov/aqs [19].

Methods

More details about the linkage of NHIS and AQS is described by Parker et al [20]. A total of 72,279 1999-2005 NHIS child survey respondents aged 3–17 years of age provided complete survey data and were eligible to be linked to pollution monitoring data. Using the longitude and latitude coordinates of the EPA monitors, air pollution exposures for $PM_{2.5}$ (quarterly weighted), PM_{10} (quarterly weighted), summer O_3, SO_2, and NO_2 were calculated from monitors within a 20-mile radius of the child's residential census block group. The number of children who were within 20 miles of a pollutant monitor varied by pollutant. Each pollutant was averaged using inverse-distance weighting and were the primary exposure measures. The outcomes for the study were based on NHIS questions "During the past 12 months, has [child's name] had any of the following conditions? Hay fever? Any kind of respiratory allergy? Any kind of food or digestive allergy? Eczema or any kind of skin allergy?" Those children whose parents reported either hay fever or respiratory allergy in the previous 12 months were classified as having respiratory allergies because both conditions result in symptoms of allergic rhinitis.

Several individual-level characteristics collected from NHIS were thought to influence the association between the air pollution and allergies and were assessed as confounding variables, including: sociodemographic characteristics, access to health services, current asthma, and adult smoking. In addition, county-level confounding variables were also considered, such as county median income from 2000 US Census, urban/rural classification, and region. Logistic regression was used to estimate odds ratios for the association between the pollutants (Summer O_3 per 10 ppb; SO_2 per 3 ppb; NO_2 per 10 ppb; PM per 10 $\mu g/m^3$) and respiratory allergy/hay fever.

Results

Researchers reported an increase in respiratory allergy/hay fever with increased summer O_3 and $PM_{2.5}$ levels. For each 10 ppb increase in O_3, the odds of child respiratory allergies or hay fever increased by 20% [adjusted odds ratio (AOR) per 10 ppb = 1.20; 95% confidence interval (CI), 1.15–1.26]. In addition for each 10 $\mu g/m^3$ increase in $PM_{2.5}$, researchers observed a 23% increase in the odds of child respiratory allergies or hay fever (AOR per 10 $\mu g/m^3$ = 1.23; 95% CI, 1.10–1.38). These observed associations remained after stratifying the analyses by urban–rural status, including multiple pollutants, and defining exposures with different exposure radii. No associations between SO_2, NO_2, and PM_{10} and the reporting respiratory allergy/hay fever were observed.

Analytic Considerations

Several limitations were identified by the researchers, and should be considered in any analyses linking geocoded data to population-based health survey data. For cross-sectional population health surveys, temporal issues of the exposure and outcome data may be problematic. The authors acknowledged that children were interviewed throughout the year, and those interviewed at the beginning of the calendar year may have been less accurately assigned exposure based on calendar-year estimates. In addition, the specific question asked about the outcome in the previous 12 months, and the timing of the outcomes may not have aligned with the timing of the exposures measured by the air monitors. While this is an issue, the authors assessed how similar the annual exposures were across the years and found that the correlations between adjacent average exposures were high (e.g., for summer O_3, r = 0.85; for $PM_{2.5}$, r = 0.80) when averaged over counties.

Secondly, geographic data collected at a single location may be inappropriately assigned over a larger administrative unit. The authors used a 20 mile radius to assign exposure to the child's census block group and recognized that a smaller radius would have been ideal, but alluded to the tradeoff between a smaller sample size and a closer air monitor versus a larger sample size and a further air monitor. The authors replicated their analysis using the subsets of children within 5 miles of air monitors and found that their results were slightly weakened for O_3 and $PM_{2.5}$, yet remained the same for other pollutants.

Furthermore, the assumption that all individuals in the larger administrative area are affected with the same levels of pollution as someone living 20 miles away from them may not be appropriate. This also leads to a variance bias trade off where the exposure estimates have lower variance but have more bias when calculated over larger, compared to smaller, areas.

3. Conclusion

Ultimately, integrating geospatial data with NCHS data systems enhances our ability to investigate broader social, economic and environmental determinants of health and disparities. It also adds value to existing data systems by providing multi-level or temporal dimensions to analyses of important public health outcomes. As compared to using geospatial data from administrative sources or complete enumerations of the population (e.g., decennial Census, vital statistics records), special considerations may be required when using geospatial data obtained from survey data (e.g., American Community Survey, Behavioral Risk Factor Surveillance System) or sampled environmental data (e.g., air pollution monitoring data from select sites). The source and type of geospatial data can present unique analytic challenges related to sampling, potential response bias, coverage and uncertainty, requiring some degree of caution when incorporating these data into analyses of public health. Researchers should be aware of the various strengths and limitations of using geospatial data in their analyses.

Disclaimer:

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

References

1. Rossen, L.M., *Neighbourhood economic deprivation explains racial/ethnic disparities in overweight and obesity among children and adolescents in the U.S.A.* J Epidemiol Community Health, 2014. **68**(2): p. 123-9.
2. Camacho-Rivera, M., et al., *Associations of neighborhood concentrated poverty, neighborhood racial/ethnic composition, and indoor allergen exposures: a cross-sectional analysis of los angeles households, 2006-2008.* J Urban Health, 2014. **91**(4): p. 661-76.
3. Heather, H.B., et al., *Epigenetics, linking social and environmental exposures to preterm birth.* Pediatric Research, 2015.

4. Juarez, P.D., et al., *The public health exposome: a population-based, exposure science approach to health disparities research.* Int J Environ Res Public Health, 2014. **11**(12): p. 12866-95.
5. Wong, W.F., T.A. LaVeist, and J.M. Sharfstein, *Achieving health equity by design.* Jama, 2015. **313**(14): p. 1417-8.
6. Gleason, J.A. and J.A. Fagliano, *Associations of daily pediatric asthma emergency department visits with air pollution in Newark, NJ: utilizing time-series and case-crossover study designs.* J Asthma, 2015. **52**(8): p. 815-22.
7. Heroux, M.E., et al., *Quantifying the health impacts of ambient air pollutants: recommendations of a WHO/Europe project.* Int J Public Health, 2015. **60**(5): p. 619-27.
8. Kioumourtzoglou, M.A., et al., *Long-term PM Exposure and Neurological Hospital Admissions in the Northeastern United States.* Environ Health Perspect, 2015.
9. Weinmayr, G., et al., *Short-term effects of PM10 and NO2 on respiratory health among children with asthma or asthma-like symptoms: a systematic review and meta-analysis.* Environ Health Perspect, 2010. **118**(4): p. 449-57.
10. Derose, K.P., et al., *Racial-Ethnic Variation in Park Use and Physical Activity in the City of Los Angeles.* J Urban Health, 2015.
11. Hankey, S., J.D. Marshall, and M. Brauer, *Health impacts of the built environment: within-urban variability in physical inactivity, air pollution, and ischemic heart disease mortality.* Environ Health Perspect, 2012. **120**(2): p. 247-53.
12. Zenk, S.N., et al., *Neighborhood retail food environment and fruit and vegetable intake in a multiethnic urban population.* Am J Health Promot, 2009. **23**(4): p. 255-64.
13. Rossen, L.M. and E.K. Kobernik, *Food insecurity and dietary intake among US youth, 2007-2010.* Pediatr Obes, 2015.
14. Curtin LR, Mohadjer LK, Dohrmann SM, et al. National Health and Nutrition Examination Survey: Sample design, 2007–2010. National Center for Health Statistics. Vital Health Stat 2(160). 2013.
15. Dugoff, E.H., M. Schuler, and E.A. Stuart, *Generalizing observational study results: applying propensity score methods to complex surveys.* Health Serv Res, 2014. **49**(1): p. 284-303.
16. National Cancer Institute, *SAS macros for Fitting Multivariate Measurement Error Models and Estimating Multivariate Usual Intake Distributions*. 2013.
17. Parker, J.D., L.J. Akinbami, and T.J. Woodruff, *Air pollution and childhood respiratory allergies in the United States.* Environmental health perspectives, 2009. **117**(1): p. 140.
18. *2004 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description*, Division of Health Interview Statistics, Editor. 2005.
19. US Environmental Protection Agency, *Air Quality System Data Mart*.
20. Parker, J., et al., *Linkage of the National Health Interview Survey to air quality data*, ed. J. Parker, et al. 2008, Hyattsville, MD: Hyattsville, MD : U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2008.